International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:10, No:8, 2016

# Air Quality Forecast Based on Principal Component Analysis-Genetic Algorithm and Back Propagation Model

Bin Mu, Site Li, Shijin Yuan

***Abstract*—**Under the circumstance of environment deterioration, people are increasingly concerned about the quality of the environment, especially air quality. As a result, it is of great value to give accurate and timely forecast of AQI (air quality index). In order to simplify influencing factors of air quality in a city, and forecast the city's AQI tomorrow, this study used MATLAB software and adopted the method of constructing a mathematic model of PCA-GABP to provide a solution. To be specific, this study firstly made principal component analysis (PCA) of influencing factors of AQI tomorrow including aspects of weather, industry waste gas and IAQI data today. Then, we used the back propagation neural network model (BP), which is optimized by genetic algorithm (GA), to give forecast of AQI tomorrow. In order to verify validity and accuracy of PCA-GABP model's forecast capability. The study uses two statistical indices to evaluate AQI forecast results (normalized mean square error and fractional bias). Eventually, this study reduces mean square error by optimizing individual gene structure in genetic algorithm and adjusting the parameters of back propagation model. To conclude, the performance of the model to forecast AQI is comparatively convincing and the model is expected to take positive effect in AQI forecast in the future.

***Keywords*—**AQI forecast, principal component analysis, genetic algorithm, back propagation neural network model.

## I. Introduction

WITH the rapid development of urbanization in China, there is an obvious rise in the living conditions of citizens, and an awareness of improving the surrounding environment also increases. Correspondingly, AQI is an important index to reflect urban air environmental quality.

This research selects Taicang in Jiangsu province, China as the research subject. Taicang city is located in the Yangtze River basin, which is close to cities with rapid economic development and a powerful industrial base. Two air monitoring stations in the urban areas of Taicang provide the data source for air quality monitoring and forecasting.

The goal of the research is to forecast Taicang City's AQI tomorrow based on air influencing factors covering aspects of weather, industry waste gas and IAQI data today.

In the data acquisition and usage, this study gathers weather, industry waste gas and individual air quality index (IAQI) of Taicang in 2015. Then, this study treats the data from the January to June as a model training set, and data from July to September as a model testing set. Afterwards, this study uses MATLAB software for simplified analysis of multi-dimensional data to decrease the dimension of the input data. Last but not least, the study input data obtained on the previous step into neural network which is optimized by genetic algorithm and draws a reasonable estimate of Taicang's tomorrow AQI value.

Bin Mu, Site Li, and Shijin Yuan are with the School of Software Engineering, Tongji University, Shanghai, China (e-mail: binmu @tongji.edu.cn, 8lisite@tongji.edu.cn, yuanshijin@tongji.edu.cn).

## II. Related Works

### *A. ANFIS Models for Air Quality Forecasting and Input Optimization*

The goal of the study is to develop an adaptive neuro-fuzzy inference system to forecast air pollution concentrations of five air pollutants in an Indian city. With the rapid development of economy and increasingly worse environment conditions, people are aware of the importance of observing, forecasting and controlling air pollution. As a result, ANFIS models is essential for building IF-THEN rules. The ANFIS model predictor considers the value of meteorological factors (pressure, temperature, relative humidity, dew point, visibility, wind speed, and precipitation) and previous day's pollutant concentration to form different combinations as the inputs to predict today's and tomorrow's air pollution concentration. The statistics for model development include the concentration value of five air pollutants and seven meteorological parameters of the city during the period 2009 to 2011. Further, the study conducted collinearity tests to eliminate the redundant input variables and a forward selection method is used for selecting the different subsets of input variables. These two methods effectively reduced the computational cost and time. Last but not least, the study evaluated the performances of the models on the basis of four statistical indices (coefficient of determination, normalized mean square error, index of agreement, and fractional bias) [1].

### *B. Application of Bias Adjustment Techniques to Improve Air Quality Forecasts*

The study includes two bias adjustment techniques; the hybrid forecast (HF) and the Kalman filter (KF). Via these techniques, the methods have been applied to investigate their capability to improve the accuracy of predictions supplied by an air quality forecast system (AQFS). The AQFS then predicts NO2, ozone, particulate matter and other pollutants

concentrations for the Lazio Region (Central Italy). A thorough evaluation of the AQFS and the two techniques has been performed through calculation, analysis of statistical parameters and skill scores. What is more, the evaluation performed considering evidenced better results for KF than for HF. RMSE scores were reduced by 43.8% (33.5% HF), 25.2% (13.2% HF) and 41.6% (39.7% HF). The statistics are respectively for hourly averaged NO2, hourly averaged O3 and daily averaged PM10 concentrations in all Lazio region monitoring sites. Eventually, a further analysis performed clustering the monitoring stations per type showed a good performance of the AQFS. The skill scores confirmed the capability of the adopted techniques to improve the reproduction of exceeding events [2].

*C. Online-Coupled Meteorology-Chemistry Model for Real-Time Air Quality Forecasting*

The study is based on an online-coupled meteorology-chemistry model called WRF/Chem-MADRID. It has been deployed for real time air quality forecast (RT-AQF) in southeastern U.S. since 2009. The model shows overall good skills in reproducing the observed multi-year trends. Besides, it is good at reproducing inter-seasonal variability in meteorological and radiative variables such as T2, WS10, Precipitation, SWDOWN, and LWDOWN, and relatively well in reproducing the observed trends in surface O3 and PM2.5. But, the study is relatively poor in reproducing the observed column abundances of CO, NO2, SO2, HCHO, TOR, and AOD. The sensitivity simulations use satellite-constrained boundary conditions for O3 and CO to show substantial improvement for both spatial distribution and domain-mean performance statistics. Last but not least, the model's forecasting skills for air quality can be further improved through improving model inputs (e.g., anthropogenic emissions for urban areas and upper boundary conditions of chemical species), meteorological forecasts (e.g., WS10, Precipitation) and meteorologically-dependent emissions (e.g., biogenic and wildfire emissions), and model physics and chemical treatments [3].

## III. Problem Discription

*A. Step Decomposition Discription*

To deal with the task more effectively, this study will solve the task of AQI forecast by step decomposition. Problems to be studied in various stages are as follows:

The main steps of this study are divided into two steps: the first is to make a simplification analysis of the influence factors of air quality, so that the data complexity can be greatly reduced, and the operation of the next step is simplified. The second is to use the data obtained in the previous step combined with the GA-BP model to give a solution of predicting Taicang's AQI value tomorrow.

*B. Symbol Description*

According to the AQI formula, this study makes a numerical simplification analysis of Taicang's AQI influencing factors, which include sulfur dioxide, nitrogen dioxide, respirable particulate matter (PM2.5 and PM10), ozone, wind speed, wind direction, precipitation and industry waste gas.

TABLE I
Symbol Description

| Symbols | Definition |
|---|---|
| $SO_2$ | Today's 24 hours average concentration of $SO_2$ ( $\mu g / m^3$ ) |
| $NO_2$ | Today's 24 hours average concentration of $NO_2$ ( $\mu g / m^3$ ) |
| $PM2.5$ | Today's 24 hours average concentration of particulate matter (particle size less than or equal to 2.5)( $\mu g / m^3$ ) |
| $PM10$ | Today's 24 hours average concentration of particulate matter (particle size less than or equal to 10) ( $\mu g / m^3$ ) |
| $O_3$ | Today's 24 hours average concentration of ozone( $\mu g / m^3$ ) |
| $AQI$ | Tomorrow's air quality index |

## IV. Model Assumptions

1. The subdivision index of urban air quality including: sulfur dioxide, nitrogen dioxide, respirable particulate matter (PM2.5 and PM10) and ozone;
2. The air quality data of Taicang air quality monitoring station and science and education city station can effectively reflect the air quality of Taicang city;
3. The weather influence factors of the air quality index mainly include wind, wind direction and precipitation;
4. The weather forecast of Taicang meteorological bureau is correct and accurate;
5. The main manual influence factor of air quality index is the waste gas emission from the surrounding factories; and,
6. The geographical distribution and the corresponding wind direction of the main waste gas emission enterprises and power plants in Taicang city have an important influence on the numerical value of AQI.

## V. Principal Compoent Analysis of Raw Data

*A. Problem Analysis*

According to the regulations of the People's Republic of China on the environmental air quality index, the AQI formula is as follows:

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + IAQI_{Lo} \quad (1)$$

$$AQI = \max\{IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n\} \quad (2)$$

Among them, $IAQI_P$ is referred to the P pollutants from the air quality index, $C_P$ is referred to the P pollutants concentration value, $BP_{Hi}$ is referred to the pollutant concentration scale value which is slightly higher than the $C_P$. $BP_{Lo}$ is referred to the pollutant concentration scale value which is a little lower than $C_P$. $IAQI_{Hi}$ corresponds to $BP_{Hi}$ while $IAQI_{Lo}$ corresponds to $BP_{Lo}$.

By analyzing AQI calculating formula, sulfur dioxide, nitrogen dioxide, particulate matter (PM2.5 and PM10), ozone and other sub indicators for AQI values have an important impact. In addition, the AQI value is also affected by wind, wind

direction, precipitation and other weather factors. What is more, the impact of the man-made factors includes factory waste gas emissions. With the help of MATLAB software, the study uses mathematical methods to screen the linear correlation of factors and gets a set of linear independent evaluation data [4]. Furthermore, this study uses the mathematics method of principal component analysis for the next step, function fitting forecast, which is prepared to reduce the overhead [5].

*B. Principal Component Analysis (PCA) Theoretical Method*

1.Basic Idea of Principal Component Analysis

Principal component analysis (PCA) is a method of linear simplification of the original multidimensional data from a mathematical point of view. The core idea of this method is to transform the original large number of variables which have a certain degree of correlation to a new generation of independent variables. In general, the mathematical method is to make a linear combination of the original variables to form a new combination of variables. It should be noted that this combination can be a lot of possibilities if it is not restricted by the corresponding conditions. If the first linear combination of the selected is the first comprehensive variable denoted as $A_1$, this study wants $A_1$ as much as possible to reflect the information of the original variable. We measure the information of variables by variance. The study hopes that Var ($A_1$) as large as possible, it is to say A contains information as much as possible. We use variance to measure the information of variables, hoping that Var ($A_1$) as far as possible, we call $A_1$ the first principal component. If the first principal component is not sufficient to represent all of the information in the original set of variables in an acceptable degree, we shall consider adding $A_2$ as second linear combinations to reflect the original information. $A_2$ does not need to reflect the information has been reflected in $A_1$, which means Cov ($A_1$, $A_2$) =0, said $A_2$ is the second principal component [6]. By this method, we can construct the analogy of third, fourth or more principal components.

2.Principal Component Analysis (PCA) Procedure

*a)* Make the original data standardized, and produce the matrix SA.

*b)* Calculate the correlation coefficient matrix R of the normalized matrix SA.

*c)* The eigenvalues and corresponding eigenvectors of the correlation coefficient matrix R are calculated.

*d)* Select main important components, and write an expression of principal component contribution rate. Assuming that the principal components have been obtained from the principal component analysis, the contribution of each principal component = $\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}$.

*e)* Calculate daily air quality principal component scores, and sort them.

*C. Principal Component Analysis Implement*

1. Firstly, this study makes a standardized processing on $SO_2$ $NO_2$, PM2.5, PM10, $O_3$, wind and wind product, precipitation, discharge amount of industry waste gas, so the numerical value obeys normal distribution. The specific operation is to collect air quality historical data, industry waste gas discharge historical data and weather conditions historical data of Taicang City in 2015 (as shown in Fig. 1, because of space limitations, the table only shows data of seven days in a year). It is essential to make pricipal component analysis of the input data for the following steps.

| Date | SO2 | NO2 | PM2.5 | PM10 | O3 | wind direction& wind product | rainfall | industry waste gas discharge | AQI |
|---|---|---|---|---|---|---|---|---|---|
| 1.01.2015 | 67 | 92 | 89 | 154 | 9 | 4 | 0 | 72 | 71 |
| 2.01.2015 | 90 | 96 | 92 | 150 | 6 | 0 | 0 | 91 | 112 |
| 3.01.2015 | 133 | 99 | 92 | 164 | 5 | -3 | 0 | 60 | 145 |
| 4.01.2015 | 173 | 101 | 101 | 161 | 5 | -1 | 0 | 53 | 198 |
| 5.01.2015 | 154 | 100 | 107 | 172 | 3 | 0 | -5 | 68 | 134 |
| 6.01.2015 | 109 | 95 | 108 | 175 | 3 | -3 | -4 | 21 | 118 |
| 7.01.2015 | 119 | 93 | 111 | 173 | 3 | -5 | 0 | 59 | 59 |

Fig. 1 AQI impact factor related data in Taicang City

It should be noted that this study has considered the relative geographical position of Taicang and Shanghai when it comes to the influence of wind direction on the air quality of Taicang. If wind is from the southeast or south, waste gas which is discharged from some industrial enterprises in Shanghai will downwind diffuse to Taicang. It will have a negative impact on air quality of Taicang. Therefore, in this study, we assign the southeast and south wind of numerical value 2, indicating the wind will greatly increase the Taicang AQI values. Accordingly, the northwest wind, north wind numerical values are assigned to -2, which indicates that the wind will greatly reduce the value of AQI in Taicang and the air quality is improved. Similarly, east wind, southwest wind was assigned to 1. West wind, the northeast wind was assigned to -1.

Because the wind power will enlarge or reduce wind impact on urban air quality, this study therefore selected the product of the direction of the corresponding numerical value and wind power as a measure of a factor of Taicang City, so as to achieve the goal of shifting from qualitative to quantitative, and combine wind factors. In addition, increase in precipitation will lead to air quality improving and AQI values decreasing, so in this study, the precipitation is considered as a factor.

Then, the columns of the data were analyzed by normal standard processing. MATLAB implementation of the standardization process is as follows:

SA(:,i)=(A(:,i)-mean(A(:,i)))/std(A(:,i))

2. Set the standardization of the data as a matrix SA, then calculate of the correlation coefficient matrix SA matrix, we can get a symmetrical matrix CM, which indicates the relationship between the I column and the J column.
3. Use MATLAB to calculate the the characteristic value D and the characteristic vector V of correlation coefficient matrix CM(D is the diagonal matrix of 8*8, its main diagonal indicates the characteristic value).

4. Descending sort of eigenvalues and then a computation of an eigenvalue of the contribution rate, then calculate from the first eigenvalue to the corresponding final eigenvalue value so far and get the accumulative contribution rate. In this study, the main component of the information retention rate is 90%, so the next step is to calculate the number of principal components, at least for how much can meet the information retention rate.
5. To obtain the feature vector PV of the principal component, with the aid of MATLAB software calculate the evaluation of the main component of the score matrix new_score through the new_score=SA*PV formula. Make a summation of the principal component score matrix according to the line, then descending sort it in the reverse arrangement. In this way, we can get nearly a year of Taicang's air quality sorted by the order of principal component score.
6. The principal component score matrix total_score can be used as the index of evaluating air quality, according to the new_score, which is obtained from the line. Together with AQI, it serves as a measure of an important index of urban air quality and this side can reflect the AQI of accuracy. Calculation of the PV feature vector group can be used as input data in step 2 and reduce the original input data dimension. It has an important role as a connecting link between the preceding and the following of forecasting of urban air quality.

## VI. The Optimization of Back Propagation Neutral Network via Genetic Algorithm

### *A. Problem Analysis*

Based on the dimension reduction data obtained by step 1, this study uses the mathematical model of neural network optimized by genetic algorithm to give a solution of predicting the future value of AQI in the city.

### *B. Air Quality Prediction Algorithm Model*

In order to improve the accuracy of prediction of air quality, this study uses a genetic algorithm to improve the BP neural network. Use genetic algorithm to optimize the initial value and threshold of the BP neural network, and then improve the accuracy of the BP neural network to predict the air quality [7].

### *C. Genetic Algorithm Theory Model*

1.Biological Basis of a Genetic Algorithm

According to Mendel's law of heredity and Darwin's theory of biological evolution, population is the basic unit in the process of biological evolution, and the essence of population evolution is the change of genetic mechanism and gene frequency [8]. Gene mutation, gene recombination, natural selection and isolation are important links in the process of species formation. Under the interaction between them, the population is gradually differentiated, and the new species is formed [9].

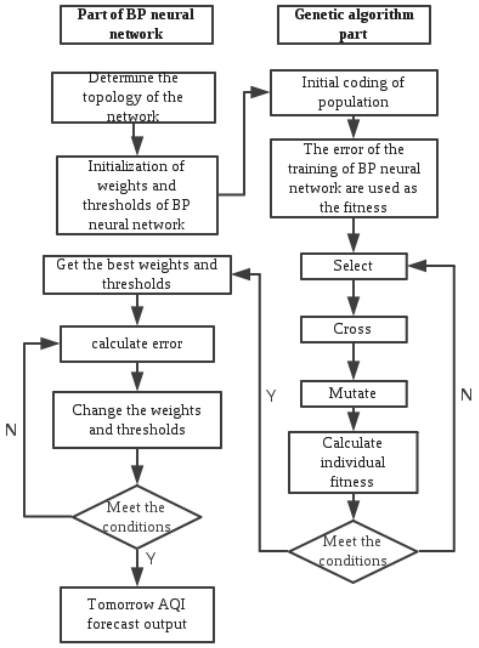


Fig. 2 Air quality prediction algorithm

2. Implementation Steps of Genetic Algorithm

*a)* *Coding and population initialization:* If a binary code of length k is used to represent the individual chromosome, it will have a different encoding of $2^k$ [10]. The chromosome encoding of each individual need to include the weights of the edge between the input layer and the hidden layer, the threshold value of the hidden layer, the value of the edge between the hidden layer and the output layer, and the threshold value of the output layer [11]. Therefore, each individual contains all the weights and thresholds of the neural network.

*b)* *Decode:* The purpose of the decoding is to restore the binary value which is not intuitive to our familiar decimal. At the macro, the genetic algorithm model of the coding and decoding operation correspond to gene and phenotype of biological. On the micro level, they correspond to the two processes of DNA transcription and translation.

*c)* *Individual fitness evaluation:* In the genetic algorithm, if the individual's adaptation is large in the current population, then the chance of individual chromosome inheritance to the next generation will be greater [12]. In the genetic algorithm to improve the neural network model, this study put the reciprocal of error between BP neural network predictive output and the expected output as the individual's adaptation value.

*d)* *Selection:* Selection operation is based on the size of the individual in the population to determine the possibility of the gene of next generation. If the total number of individuals in the population is N, the fitness of individual I is $f_i$, then the probability of individual I is selected as the

$$P_i = \frac{f_i}{\sum_{k=1}^{N} f_k} \quad (3)$$

where the probability of selecting each individual is determined, we need to produce a random number between interval [0, 1] to decide which individuals participate in replication. If the individual's fitness is high, then the selecting probability $P_i$ is high and its genetic gene is likely to spread in the population. If the individual's adaptability is small, it will be gradually eliminated.

*e) Cross:* Crossing is the selection of the chromosomes of two individuals from the population, and the corresponding parts of them are exchanged to generate new individuals. It should be noted that the location of the chromosome is randomly selected.

*f) Mutation:* Mutation is a mutation in the chromosome of a particular individual in a population to produce a new individual. For example, the chromosome S=11001101, if we change the fourth gene 0 into 1, that is 11011101.

*D. Theoretical Model of BP Neural Network*

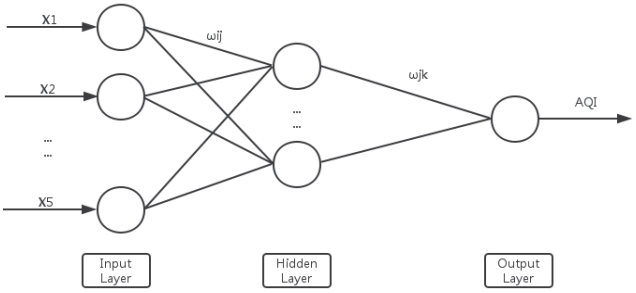


Fig. 3 Neural network topology

1. *Transfer Function*

**a. Linear Type (Used in Input Neurons and Output Neurons):**

$$f(x) = x \quad (4)$$

**b. S Type (Used in the Hidden Layer Neurons):**

$$f(x) = \frac{1}{1+e^{-x}} \quad (5)$$

In this study, the variable of the incentive function is coupled with a constant c, so that $e^{-x}$ into $e^{-\alpha}$ in order to adjust the magnitude of the transfer function.

2. *Basic Mathematical Principles of BP Neural Network*

A BP neural network is a kind of multi-layer feed forward neural network. Its name derives from the process of training in the network, and the algorithm of adjusting the weights of network is the back-propagation algorithm. Fig. 2 shows the topology of a common BP neural network.

In this paper, the BP neural network has three layers of neurons, including the input layer, hidden layer and output layer [13]. Between the upper and lower levels, there is full connectivity, while the same layer of neurons is without connection [14].

Neurons in the input layer and the hidden layer neurons between the weights of the network $\omega_{ij}$ , implicit between layer neurons and the neurons in the output layer is the weights of the network $\omega_{jk}$ . In addition, hidden layer and output layer not only go into neuronal integration [15] of all neurons in the previous layer's information, but also in the process of integration adds a threshold, which is to mimic biological neurons a certain threshold must be reached that would trigger the principle.

When the input values and output value of learning samples are provided to neurons, neuronal activation values spread from the input layer through the hidden layer to the output layer. The neurons in the output layer react to the input of the network and reduce error between the predicted output and the desired output in the direction from the output layer reversely through the hidden layer back to the input layer and modify the weights step by step.

*E. GA-BP Model Implement*

In the process of the algorithm, this study treats the data from January 2015 to June 2015 as the training set of the model, and the data from July 2015 to September 2015 as the testing set of the model. Each individual in the population contains the weights and thresholds of the whole BP neural network. Each individual calculates their adaptation values by fitness function, through selection, crossover and mutation operation to find the optimal adaptive value corresponding to the individual. BP neural network prediction uses the genetic algorithm to get the optimal individual as the initial network weights and threshold values and train the network to get the prediction function, and then to test the size of the error.

At the initial time, the BP neural network determines the structure of the BP neural network according to the input and output parameters of the fitting function. In this study, because of the five input parameters (five linearly independent vectors in step one) and one output parameter (today AQI values), according to the Kolmogorov theorem states that the hidden layer node number s = 2n + 1 (n is the input layer nodes), we can select neural network hidden layer nodes number as 2 * 5+1=11.

After the structure of the BP neural network is determined, the length of the chromosome in the genetic algorithm can be determined. Because the nonlinear fitting function has five input parameters and one output parameter and 11 intermediate nodes, this study set BP neural network input layer 5 nodes, hidden layer 11 nodes, 1 node in the output layer, total 5 x 11+11 * 1=66 weights, 11+1=12 thresholds. Therefore, in this study each individual in the genetic algorithm coding length is set to 78.

Then genetic algorithm uses the code contained in the neural network weights and thresholds and treats the reciprocal of the error between predicted output and the expected output value as the fitness function. The greater the fitting value of the individual genetic is, the greater the probability of the next generation will be. Finally, the fitness of the individual's

chromosome value as the initial weight and threshold of the BP neural network is obtained.

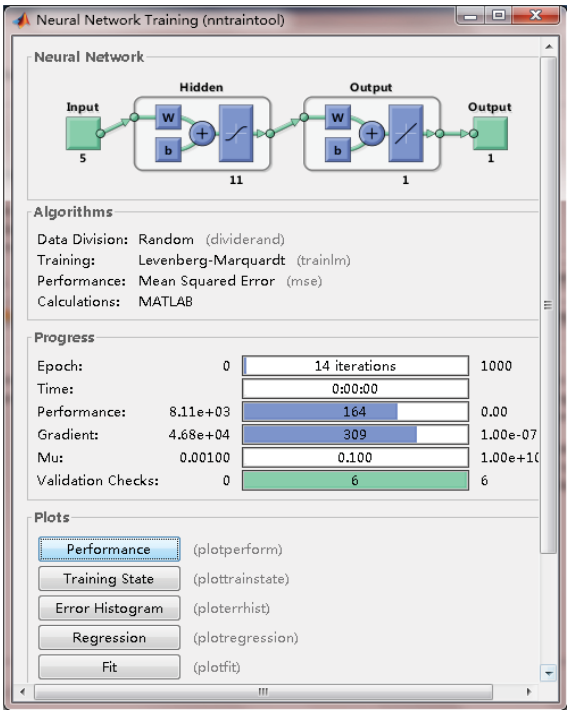


Fig. 4 Neural network toolbox

After determining the initial weights and thresholds of BP neural network, this study uses MATLAB neural network toolbox to give a fitting function of input data and output data and obtain the function relationship of the input data and tomorrow AQI. Finally, input test data to test the reliability of the model, and then use the model to predict the actual tomorrow AQI. Because the current air quality prediction is too difficult, and the influence factors are too varied, volatile and unstable, this paper choose the next 24 hours of air quality AQI forecast, to get more accurate results.

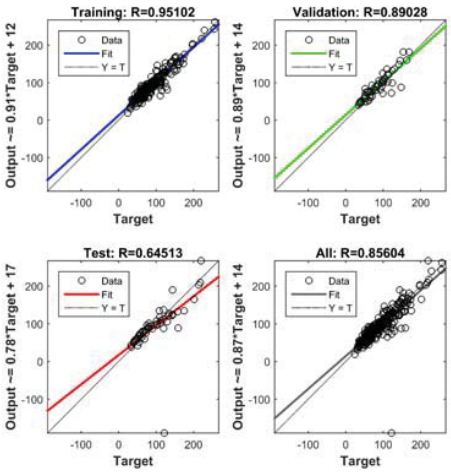


Fig. 5 Functional fitness regression diagram

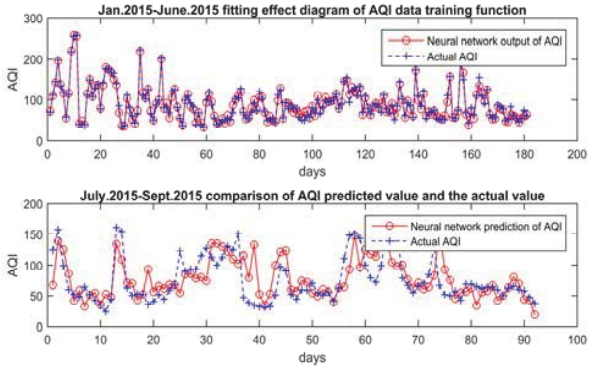


Fig. 6 Function performance test and test data prediction

## VII. Result Analysis and Model Improvement

### *A. Result Analysis and Error Analysis*

At first, the study makes a performance validation of the PCA-GABP model. It is obvious that the best validation performance is at epoch 9.

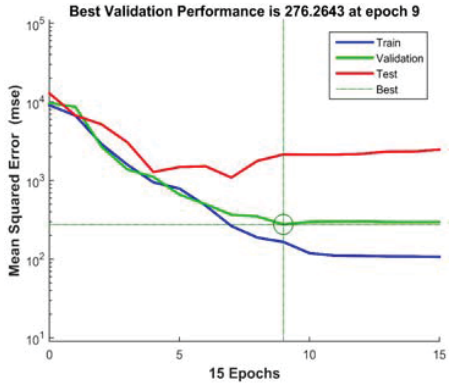


Fig. 7 Validation performance analysis

What is more, the performances of the models were evaluated on the basis of two statistical indices (normalized mean square error and fractional bias) [16].


Fig. 8 Error histogram

In contrast to previous studies, in which fractional bias of air quality predicted is around 30% to 35% [17], this study reduces

mean square error by adjusting the parameters of the PCA-GABP model through numerous experiments and finally reduces the fractional bias of prediction to 29.5% [18]. As a result, the study is comparatively convincing and of practical value to real life.

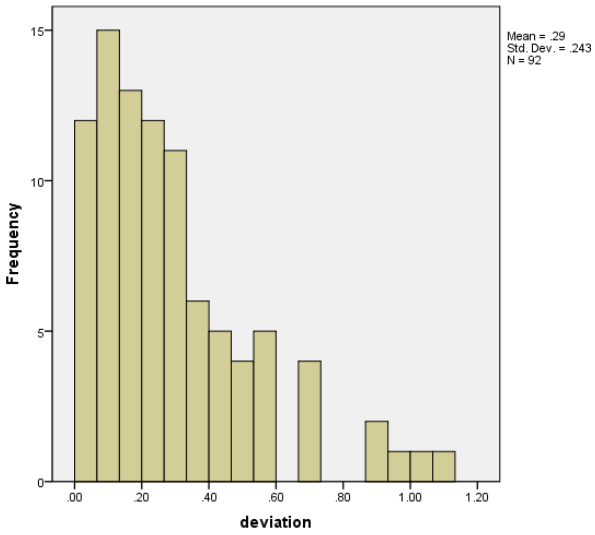


Fig. 9 Fractional bias histogram

*B. Improvement of Model*

In the implementation process of the algorithm, initial population size, crossover, and mutation probability parameter values set for the problem play a very important role for finding a satisfactory solution, to improve the value they will greatly improve the model's accuracy.

First of all, if the population size is too small, the population will inbreed and the probability of generating a competitive individual becomes small. If the population size is too large, it will cause the convergence difficulties, so this study chose a population size of 50. What is more, mutation probability is too small to make the population's diversity decline too fast. If the variation of the probability value is too large, it will cause convergence difficulties, so it is in the range of 0.0001 ~ 0.2. Cross probability is too small to ensure that the population be effectively updated. If the crossover probability value is too large, it is easy to destroy the original fine model and miss the outstanding individuals, so it is in the range of 0.4 ~ 0.99. Generation of an evolutionary population is not easy to facilitate obtaining the algorithm convergence if population evolutionary generation is too small. The algorithm has long been convergent and will waste resources if population evolutionary generation is too large, so in this study, the evolutionary generation is set to 100. In addition, although the initial population generation is random, it is important to estimate that the initial population which is close to the optimal solution, so the cost of the algorithm is reduced.

When the neural network structure is set, the input layer nodes and output layer nodes are immediately identified, and the first very important and difficult problem we then encounter is how to optimize the hidden layer nodes and hidden layers. The experiment shows that the network does not have the necessary learning ability and information processing ability if the hidden layer nodes are too small. Conversely, if too much, it will not only greatly increase network structure complexity (this point is especially important to hardware implementation of the network), but also make the learning speed of the network becomes very slow. The selection of the number of hidden nodes has always been highly valued by the researchers of neural networks. Gorman pointed out the relationship between the hidden layer nodes and mode number n: $S = \log_2 N$ ; Kolmogorov theorem shows that, the hidden layer node number s = 2n + 1 (n is the input layer nodes); and according to literature: s = sqrt (0.43mn + 0.12nn + 2.54M + 0.77n + 0.35 in the + 0.51(m is the number of input layer, n is number of output layer).

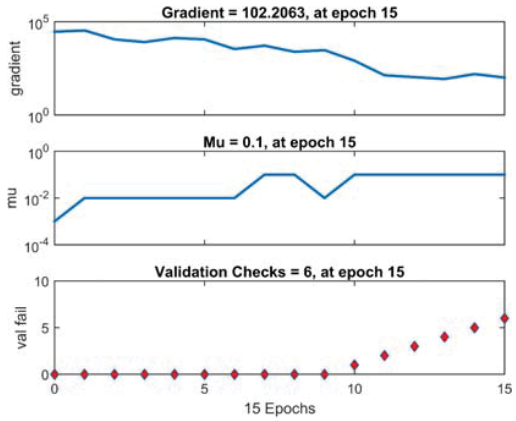


Fig. 10 Training state analysis

*C. Generalization of the Model*

The prediction of urban air quality AQI algorithm in this study has a broad application in the problem for production scheduling, machine learning, function fitting, image recognition etc., so the study and improvement the prediction of urban air quality AQI algorithm is significant to actual production and life.

## Acknowledgment

The authors thank to the sincere support from the instructors and students in their laboratory. They express that their efforts are of huge value to the survey of the experiment, and the strong support of Economy and that Information Department of Taicang, China is also essential to this research.

## References

[1] Kanchan Prasad, Amit Kumar Gorai, Pramila Goyal, Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time (J). Atmospheric Environment, 128(2016): 246-262, 2016.

[2] Camillo Silibello, Alessio D'Allura, Sandro Finardi, Application of bias adjustment techniques to improve air quality forecasts (J). Atmospheric Pollution Research, 6(6): 928–938, 2015.

[3] Yang Zhang, Chaopeng Hong, Khairunnisa Yahya, Comprehensive evaluation of multi-year real-time air quality forecasting using an online-coupled meteorology-chemistry model over southeastern United States (J). Atmospheric Environment, 138 (2016): 162–182, 2016.

[4] Zhuo Jinwu, MATLAB in the application of mathematical modeling (M). Beijing: Beihang University press, 2014.
[5] Wang Xiaochuan, Shi Feng, Yu Lei, Li Yang, MATLAB neural network 43 case analysis (M). Beijing: Beihang University press, 2013
[6] Meng Dong, fan Zhongjun, Wang Jiazhen, chaos genetic algorithm to the BP neural network improved (J). Mathematical theory and application, 34 (1): 102-110, 2014.
[7] Liu Yuanyuan, Lian Jijian, Zhu Yun, Application of improved BP neural network based on Genetic Algorithm in prediction of chaotic runoff time series(J).Hydrology, 27 (2): 45-48, 2007.
[8] Li Song, Liu Lijun, de Yongle, Chaos prediction of short time traffic flow based on genetic algorithm optimized BP neural network (J).Control and decision, 26 (10): 1581-1585, 2011.
[9] Gao Daqi, teachers of linear basis function to the three layer neural network (J). Journal of computer, 21 (1) 80-86, 1998.
[10] Park, Yang-Byung, Yoo, Jun-Su, Park, Hae-Soo, A genetic algorithm for the vendor-managed inventory routing problem with lost sales (J). Expert systems with applications, 53, 149-159, 2016.
[11] Ardjmand, Young, WA, Weckman, GR, Applying genetic algorithm to a new bi-objective stochastic model for transportation, location, and allocation of hazardous materials (J). Expert systems with applications, 51, 49-58, 2016.
[12] Contreras-Bolton, Carlos, Gatica, Gustavo, A multi-operator genetic algorithm for the generalized minimum spanning tree problem (J). Expert systems with applications, 50, 1-8, 2016.
[13] Hwang, Inyoung,, Park, Hyung-Min, Chang, Joon-Hyuk, Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection (J). Computer speech and language, 38, 1-12, 2016.
[14] Barat, Cecile, Ducottet, Christophe, String representations and distances in deep Convolutional Neural Networks for image classification (J). Pattern recognition, 54, 104-115, 2016.
[15] Li, Fanjun, Qiao, Junfei, Han, Honggui, A self-organizing cascade neural network with random weights for nonlinear system modeling (J). Applied soft computing, 42, 184-193, 2016.
[16] LiuYan, Cheng Zhi-long, Xu Jing, Improvement and Validation of Genetic Programming Symbolic Regression Technique of Silva and Applications in Deriving Heat Transfer Correlations (J). Heat transfer engineering, 37(10), 862-874, 2016.
[17] Prabhu, M. Venkatesh, Karthikeyan R, Modeling and optimization by response surface methodology and neural network-genetic algorithm for decolorization of real textile dye effluent using Pleurotus ostreatus: a comparison study (J). Desalination and water treatment, 57(28), 13005-13019, 2016.
[18] Khoshbin, Fatemeh, Bonakdari, Hossein, Adaptive neuro-fuzzy inference system multi-objective optimization using the genetic algorithm/singular value decomposition method for modelling the discharge coefficient in rectangular sharp-crested side weirs (J) Engineering optimization, 48(6), 933-948.